iScience

CellPress

Special Issue: RECOMB-Seq 2019

Editorial

Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going?

A wave of technologies transformed sequencing over a decade ago into the high-throughput era, demanding research in new computational methods to analyze these data. The applications of these sequencing technologies have continuously expanded since then. The RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-Seq) meeting, established in 2011, brings together leading researchers in computational genomics and genomic biology to discuss emerging frontiers in algorithm development for massively parallel sequencing data. The ninth edition of this workshop was held in Washington, DC, in George Washington University on May 3 and 4, 2019. There was an exploration of several traditional topics in sequence analysis, including genome assembly, sequence alignment, and data compression, and development of methods for new sequencing technologies, including linked reads and single-molecule long-read sequencing. Here we revisit these topics and discuss the current status and perspectives of sequencing technologies and analyses.

Advances in high-throughput sequencing technologies provide holistic investigatory capabilities to address critically important and complex problems in virtually every area of biology. These technologies have led to an explosive growth of the amount of sequencing data being generated every year. For example, the Human Genome Project cost billions of dollars and took a decade to complete, whereas more than 100,000 human genomes have been sequenced over the past 5 years. In addition to the advancements in throughput and cost, a number of novel sequencing technologies have emerged, including ultra-long read sequencing (e.g., Nanopore), high-resolution restriction maps (e.g., Bionano data), linked-read sequencing technologies, and cross-linking methods (see Table 1 for an overview of various sequencing technologies). All these sequencing technologies have been a source of discussion and intense research. Here, we summarize some of the most recent findings and opportunities.

Assembly of genomes using shotgun sequencing of chromosomal DNA still remains a fundamental problem in the bioinformatics community. As stated by Adam Phillippy in his talk "40 Years of Genome Assembly: Are We Done Yet?" the development of strategies to assemble sequence reads was described in the late 1970s when (Staden, 1979) stated that "With modern fast sequencing technologies and suitable computer programs it is now possible to sequence whole genomes without the need of restriction maps." Four decades later, although reference genomes have been assembled for a number of organisms, including humans, significant challenges remain in obtaining complete assemblies routinely. As described by Phillippy, the current human reference genome (GRCh38) contains 102 gaps and lacks sequence for centromeres and other repetitive regions. Phillippy described the first "telomere-to-telomere" (T2T) assembly of the human chromosome X using a combination of long (e.g., PacBio technology) and ultra-long Nanopore sequence reads. This notable success could lead to the generation of a T2T assembly of all human chromosomes in the near future. The challenges that remain in this area are not new, but they have continued to haunt researchers for many years, namely, long repeats, heterozygosity, data accuracy, and measuring assembly quality.

Toward overcoming these challenges, there have been a number of proposed computational solutions that improve the quality of both the assembly and the sequence data. Morisse et al. (2019) proposed a method for self-correction of long-read data, which combines algorithmic approaches of current state-of-the-art long-read error correction methods, namely, construction and use of multiple alignment of the reads, and subsequently, a de Bruijn graph. They demonstrate that the method is able to error correct both long and ultra-long sequence reads and is highly scalable, as it is the only method that is able to scale to a human dataset containing ultra-long reads. Marijon et al. (2019) describe a method to analyze assembly graphs produced from long reads to recover contigs that were lost during the assembly process. They

Technology	Method	Read Length	Error Rate (%)	Throughput (GB/run)
Illumina	Synthesis	100–300 bp	0.1	200–600
Pacific Biosciences SMRT	Synthesis	10–100 kb	5–15	10–20
Oxford Nanopore MinION	Nanopore	Variable (up to 1,000 kb)	5–20	5–10

Table 1. Comparison of the Read Lengths, Error Rates, and Costs of Various DNA Sequencing Technologies

demonstrate that their method recovers useful adjacency information between contigs and show that it is able to "provide a more informative representation of fragmented assemblies, examine repeat structures, and propose likely contig orderings." In a similar spirit Shlemov and Korobeynikov (2019) develop a method for analyzing the assembly graph by aligning a profile hidden Markov models to the graph to discover the set of most probable paths in the graph. It is suggested that this information can be used for putative gene finding in metagenomic samples, repeat resolution, or scaffolding. These works suggest that there is significant opportunity to improve upon error correction and assembly of long-read sequencing data and a surge in interest and potential use of ultra-long sequencing in genome assembly. Last, large sequencing projects, such as the Vertebrate Genome Project, foreshadow the need for hybrid assembly approaches and assembly frameworks wherein algorithmic ideas and approaches can be easily validated.

For species with an assembled genome, the sequence reads can be aligned to this reference genome to identify genetic variants and perform a variety of other biological analyses. Therefore, alignment of DNA sequences to a genome is a fundamental computational problem. A number of methods have been developed for the problem of aligning short reads (50–200 bases in length) to a reference genome over the past 10 years (Reinert et al., 2015). Many of these alignment tools have been developed specifically for aligning reads generated using next-generation sequencing (NGS) protocols such as RNA sequencing (RNA-seq) and microRNA (miRNA) sequencing. Reads generated using RNA-seq can span exon-exon junctions, and therefore accurate mapping of RNA-seq reads requires the ability to detect spliced alignments. Zhong and Zhang (2019) described an alignment tool designed to enable the accurate mapping of cross-linked miRNA-mRNA reads. This tool uses a Burrows-Wheeler Transform (BWT)-based index for finding short matches but implements a number of additional optimizations to enable the sensitive mapping of duplex reads formed by miRNA-mRNA interactions. Compared with existing alignment tools such as STAR and BLASTN, this specialized alignment tool, CLAN, maps more reads and has greater accuracy.

With the emergence of single-molecule long-read sequencing technologies such as Pacific Biosciences SMRT and Oxford Nanopore MinION (Pollard et al., 2018), there is an increasing need for alignment tools capable of aligning long reads. Existing NGS alignment tools are optimized for low error rates and short read lengths, whereas these technologies generate reads that are tens of kilobases long and have high error rates (5%–20%, see Table 1). Almost all short-read alignment tools use a hash table or a BWT-based index to efficiently find short matches between a query sequence and a genome. Hash-table-based approaches require the storage of a large index for finding the seed matches, which can be space prohibitive for large genomes such as those of humans. Li described the use of "minimizers," an elegant idea that enables the detection of seed matches while storing only a fraction of the seeds (Roberts et al., 2004), to design a long-read alignment tool, Minimap2 (Li, 2018). This tool combines the use of minimizers with chaining and affine gap alignment to efficiently align both long DNA reads and cDNA/mRNA reads.

Detection of genetic variants using sequence reads aligned to a reference genome is perhaps the most common application of NGS technologies. Similar to read alignment, many tools have been developed to detect short sequence variants (single nucleotide variants and short insertions or deletions). Using state-of-the-art tools such as GATK (DePristo et al., 2011) both these types of variants can be reliably detected using whole-genome or whole-exome DNA sequencing. Nevertheless, other types of variants such as structural variants remain challenging to detect using NGS reads. Melissa Gymrek highlighted one such limitation of NGS for short tandem repeats (STRs). STRs (tandem repeats of 1- to 6-base-long motifs) are abundant in the human genome and are prone to mutations that can expand or contract the repeat.

Expansions of STRs have been shown to cause a number of rare Mendelian diseases (Ashley, 2016). One example of such a disease is Huntington disease, which is caused by the expansion of a trinucleotide repeat. Genotyping of STRs and detection of repeat expansions requires careful analysis to capture the signal for such events in short sequence reads. Gymrek described a computational tool, GangSTR (Mousavi et al., 2019), that can accurately genotype STRs at more than 500,000 tandem repeat loci and even detect repeat expansions that are longer than the length of Illumina reads. Nevertheless, many challenges remain in this area, including genotyping GC-rich repeats and accounting for nonuniformity in sequence coverage.

Similar to read mapping, detection of variants in different applications (e.g., somatic variants in cancer genomes) requires specialized tools. Charlotte Darby presented a clever approach (Darby et al., 2019) to detect mosaic variants using the 10X Genomics linked-read technology (Zheng et al., 2016). Unlike germline variants, mosaic variants are those that are present in only a subset of the cells of an individual and are harder to detect. In contrast with standard Illumina sequencing, linked reads provide long-range haplotype information that can be leveraged for discriminating mosaic mutations (present on a subset of reads from one haplotype) from sequencing errors and other artifacts. The novel method, Samovar, assigns reads to haplotypes using the linked reads and enables accurate detection of mosaic mutations in pediatric cancer genomes without the use of matching normal datasets.

Tools such as GangSTR and Samovar are crucial for realizing the full potential of whole-genome sequencing and will further enhance the use of NGS as a diagnostic tool. One limitation of these tools is that they rely on the alignment of reads to a reference genome and are designed to detect specific types of variants. There is a growing interest in alignment-free variant detection and genotyping methods for NGS data. Such methods utilize the information contained within the set of k-mers (and their counts) observed in the sequence reads and can be used to detect almost all types of sequence variants (Nordstrom et al., 2013). Daniel Standage presented a method, Kevlar (Standage et al., 2019), that detects *de novo* variants in an individual's genome by identifying frequent k-mers that are either completely absent or appear at very frequency in the genomes of the parents. This alignment-free approach can detect SNVs, short indels, and even structural variants. Alignment-free approaches are also valuable for genotyping known variants in a sequenced genome. Luca Denti described MALVA (Bernardini et al., 2019), which can genotype both SNVs and short indels efficiently and improves upon previous methods for this problem. The success of alignment-free methods suggests that approaches that combine k-mer-based analysis with reference-based mapping could maximize accuracy for variant detection and genotyping using NGS reads.

The 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) is now largely completed, and now the 100,000 Genomes Project is well underway (Turnbull et al., 2018). With no compression, the raw data for 100,000 human genomes requires roughly 300 terabytes of disk space. Given the size of the data and its continual growth, efficient compression and decompression of the data is vital to any sort of analysis. There are different methods to tackle data compression, which is frequently but not necessarily reliant on the analysis goals. General compression algorithms, such as Lempel-Ziv parsing (Ziv and Lempel, 1977), BWT (Burrows and Wheeler, 1994), and Huffman encoding (Turnbull et al., 2018), aim to transform the input file(s) into a representation that requires fewer bits than the original file(s). Conversely, decompression aims to recover the original files from the compressed format.

Sequence data offers unique opportunities for significant compression because it contains high levels of redundancy. A number of methods that utilize novel data structures to exploit this characteristic to achieve efficient compression and decompression have been developed. Sequence Bloom Trees and space-efficient de Bruijn graph representations are two examples of such data structures that have been continuously improved upon in the past few years. Sequence Bloom Trees were first proposed by Solomon and Kingsford (2016) as a means to efficiently index sequence data in a manner that supports queries about the presence of transcripts. SeqOthello (Yu et al., 2018), Split-SBT (Solomon and Kingsford, 2018), and AllSome-SBT (Sun et al., 2017) improve upon this recent original representation. Paul Medvedev described a representation (Harris and Medvedev, 2019) that requires substantially less time and space to construct the index, demonstrating that there remains opportunity for further improvements to existing representations. Comparably, de Bruijn graphs, which were originally proposed for genome assembly, have been used to compactly index all *k*-length subsequences (*k*-mers) from a set of sequence reads. Although there have been numerous improvements in the representation of de Bruijn graphs (Muggli et al., 2017;

Almodaresi et al., 2019; Karasikov et al., 2019; Almodaresi et al., 2017; Alipanahi et al., 2018; Mustafa et al., 2017; Pandey et al., 2018), we still continue to witness substantial improvements on existing representations. For example, the representation of Marchet et al. (2019) was able to index all *k*-mers from the human genome in 8-GB space and 30 min and all *k*-mers from the *axolotl* genome (10 times the size of the human genome) in 63-GB space and within 10 h. This area of using de Bruijn graphs for compactly representing and indexing *k*-mers still has unexplored avenues.

Last, there is still significant work in developing targeted parallelism to rapidly compress and decompress large gzip files. Kerbiriou and Rayan presented a parallel algorithm for fast decompression of gzip-compressed files that allows random access to compressed DNA sequence data in the FASTQ format. Demonstrations of their method show that it is an order of magnitude faster than gunzip, and five times faster than a highly optimized sequential implementation.

Two recurring themes emerged at RECOMB-Seq 2019. First, although computational methods for alignment, assembly, and variant detection have advanced tremendously over the past decade, significant challenges remain, e.g., end-to-end genome assembly using long reads and detection of repeat variants using high-throughput sequencing. Second, there is a need for new algorithms and data structures to process data generated from multiple genomes and using newer sequencing technologies. In particular, long-read sequencing technologies such as Pacific Biosciences SMRT and 10X Genomics linked-read sequencing are becoming increasingly ubiquitous and we expect to see the development of new methods that leverage these technologies in the near future. Last, the meeting would not have been a success without the diligent work of the members of the program and steering committees. We would like to thank everyone who contributed to making RECOMB-Seq 2019 a success.

ACKNOWLEDGMENTS

C.B. was funded by the NIH through NIAID grant R01AI141810-01 and by the NSF through grant IIS-1618814.

Vikas Bansal[1]
Christina Boucher[2]
[1]Department of Pediatrics, School of Medicine, University of California, San Diego, La Jolla, CA, USA
[2]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA

https://doi.org/10.1016/j.isci.2019.06.035

REFERENCES

Alipanahi, B., Kuhnle, A., and Boucher, C. (2018). Recoloring the colored de Bruijn graph. In: Proceedings of String Processing and Information Retrieval (SPIRE 2018), pp. 1–11.

Almodaresi, F., Pandey, P., and Patro, R. (2017). Rainbowfish: a succinct colored de Bruijn graph representation. In: Proceedings of the Workshop of Algorithms in Bioinformatics (WABI 2017), pp. 251–265.

Almodaresi, F., Pandey, P., Ferdman, M., Johnson, R., and Patro, R. (2019). An efficient, scalable and exact representation of high-dimensional color information enabled via de Bruijn graph search. In: Proceedings of the Research in Computational Molecular Biology (RECOMB 2019), pp. 1–18.

Ashley, E.A. (2016). Towards precision medicine. Nat. Rev. Genet. *17*, 507–522.

Bernardini, G., Bonizzoni, P., Denti, L., Previtali, M., and Alexander Schönhuth, A. (2019). MALVA: genotyping by Mapping-free ALlele detection of known VAriants. BioRxiv. https://doi.org/10.1101/575126.

Burrows, M. and Wheeler, D.J. (1994). A block sorting lossless data compression algorithm, Technical Report 124, Digital Equipment Corporation, https://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.pdf.

Darby, C.A., Fitch, J.R., Brennan, P.J., Kelly, B.J., Bir, N., Magrini, V., Leonard, J., Cottrell, C.E., Gastier-Foster, J.M., Wilson, R.K., et al. (2019). Samovar: single-sample mosaic SNV calling with linked reads. iScience. https://doi.org/10.1016/j.isci.2019.05.037.

DePristo, M.A., Quintero, J.C., Cruz, D.F., Quintero, C., Hubmann, G., Foulquié-Moreno, M.R., Verstrepen, K.J., Thevelein, J.M., and Tohme, J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Harris, R.S., and Medvedev, P. (2019). Improved representation of sequence Bloom trees. BioRxiv. https://doi.org/10.1101/501452.

Karasikov, M., Mustafa, H., Joudaki, A., Javadzadeh-No, S., Rätsch, G., and Kahles A. (2019). Sparse binary relation representations for genome graph annotation. In: Proceedings of the Research in Computational Molecular Biology (RECOMB 2019). pp. 120–135. https://doi.org/10.1101/468512.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Marchet, C., Kerbiriou, M., and Limasset, A. (2019). Indexing de Bruijn graphs with minimizers. BioRxiv. https://doi.org/10.1101/546309.

Marijon, P., Chikhi, R., and Varré, J.S. (2019). Graph analysis of fragmented long-read bacterial genome assemblies. Bioinformatics, btz219, https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz219/5421164?redirectedFrom=fulltext.

Morisse, P., Marchet, C., Limasset, A., Lecroq, T., and Lefebvre, A. (2019). CONSENT: scalable self-correction of long reads with multiple sequence alignment. BioRxiv. https://doi.org/10.1101/546630.

Mousavi, N., Shleizer-Burko, S., Yanicky, R., and Gymrek, M. (2019). Profiling the genome-wide

landscape of tandem repeat expansions. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz501.

Muggli, M.D., Bowe, A., Noyes, N.R., Morley, P.S., Belk, K.E., Raymond, R., Gagie, T., Puglisi, S.J., and Boucher, C. (2017). Succinct colored de Bruijn graphs. Bioinformatics *33*, 3181–3187.

Mustafa, H., Kahles, A., Karasikov, M., and Raetsch, G. (2017). Metannot: a succinct data structure for compression of colors in dynamic de Bruijn graphs. BioRxiv. https://doi.org/10.3929/ethz-b-000236153.

Nordstrom, K.J., Albani, M.C., James, G.V., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G., and Schneeberger, K. (2013). Mutation identification by direct comparison of whole- genome sequencing data from mutant and wild-type individuals using k-mers. Nat. Biotechnol. *31*, 325–330.

Pandey, P., Almodaresi, F., Bender, M.A., Ferdman, M., Johnson, R., and Patro, R. (2018). Mantis: a fast, small, and exact large-scale sequence-search index. Cell *7*, 201–207.

Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., and Sandhu, M.S. (2018). Long reads: their purpose and place. Hum. Mol. Genet. *27* (R2), R234–R241.

Reinert, K., Langmead, B., Weese, D., and Evers, D.J. (2015). Alignment of next-generation sequencing reads. Annu. Rev. Genomics Hum. Genet. *16*, 133–151.

Roberts, M., Hayes, W., Hunt, B.R., Mount, S.M., and Yorke, J.A. (2004). Reducing storage requirements for biological sequence comparison. Bioinformatics *20*, 3363–3369.

Shlemov, A., and Korobeynikov, A. (2019). PathRacer: racing profile HMM paths on assembly graph. BioRxiv. https://doi.org/10.1101/562579.

Solomon, B., and Kingsford, C. (2016). Fast search of thousands of short-read sequencing experiments. Nat. Biotechnol. https://doi.org/10.1038/nbt.3442.

Solomon, B., and Kingsford, C. (2018). Improved search of large transcriptomic sequencing databases using split sequence bloom trees. J. Comput. Biol. *25*, 755–765.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. *6*, 2601–2610.

Standage, D.S., Titus Brown, C., and Hormozdiari, F. (2019). Kevlar: a mapping-free framework for accurate discovery of de novo variants. BioRxiv. https://doi.org/10.1101/549154.

Sun, C., Harris, R.S., Chikhi, R., and Medvedev P. (2017). AllSome sequence bloom trees. In: Proceedings of the Research in Computational Molecular Biology (RECOMB 2017), pp. 272–286, https://doi.org/10.1007/978-3-319-56970-3_17.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Turnbull, C., Scott, R.H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F.B., Halai, D., Baple, E., Craig, C., Hamblin, A., et al. (2018). The 100,000 Genomes Project: bringing whole genome sequencing to the NHS. Br. Med. J. *361*, k1687.

Yu, Y., Liu, J., Liu, X., Zhang, Y., Magner, E., Lehnert, E., Qian, C., and Liu, J. (2018). SeqOthello: querying RNA-seq experiments at scale. Genome Biol. *19*, 167.

Zheng, G.X., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. *34*, 303–311.

Zhong, C., and Zhang, S. (2016). CLAN: the CrossLinked reads ANalyais tool. iScience. https://doi.org/10.1016/j.isci.2019.05.038.

Ziv, J., and Lempel, A. (1977). A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory *23*, 337–343.